

OBSERVING ONLINE DICTIONARY USERS: STUDIES USING WIKTIONARY LOG FILES

Carolín Müller-Spitzer: *Institute for the German Language, Mannheim*
(mueller-spitzer@ids-mannheim.de)

Sascha Wolfer: *Institute for the German Language, Mannheim*
(wolfer@ids-mannheim.de)

Alexander Koplenig: *Institute for the German Language, Mannheim*
(koplenig@ids-mannheim.de)

Abstract

We present studies using the 2013 log files from the German version of Wiktionary. We investigate several lexicographically relevant variables and their effect on look-up frequency: Corpus frequency of the headword seems to have a strong effect on the number of visits to a Wiktionary entry. We then consider the question of whether polysemic words are looked up more often than monosemic ones. Here, we also have to take into account that polysemic words are more frequent in most languages. Finally, we present a technique to investigate the time-course of look-up behaviour for specific entries. We exemplify the method by investigating influences of (temporary) social relevance of specific headwords.

1. Introduction¹

Observing dictionary users via log file studies is a promising method for research into dictionary use (e.g., Bergenholtz & Johnsen 2005, De Schryver et al. 2006, De Schryver 2013a, De Schryver & Joffe 2004, Lew 2011a), since it enables the researcher to record real behaviour of dictionary users in a natural setting. However, the method is limited because the researcher – as is the case for all observing methods – has no control over the research process. In other words, it is hardly possible to find out anything about the background of the observed users, the contexts of dictionary use, the success of the look-up process etc. (Bergenholtz & Johnsen 2013: 558, cf. also Hult 2012: 924, Tarp 2009: 289–290, Tiberius & Niestadt 2014). Therefore, Lew urges not to draw too many conclusions from individual look-up processes recorded in log files,

since “we typically know precious little about the user” and one “cannot be sure that the user has selected an even remotely appropriate tool for the job”. For this reason, we should not “modify the dictionary to dutifully and indiscriminately serve all types of oddball queries” (Lew 2011a: 7). Correspondingly, one for example has to be careful not to interpret look-up tasks performed by one IP address as look-up tasks by one single dictionary user, because one single IP-address does not necessarily correspond to one dictionary user (cf. Santos & Frankenberg-Garcia 2007: 356–359).

Combining look-up data with data gained from other methods of empirical research (e.g., a post look-up questionnaire) to enrich log file data later on with information about personal background, user needs etc. (using IP addresses to merge the datasets), as Hult proposes (Hult 2012: 924–926), is – at least in Germany – prohibited without explicit user permission.² At that point, the user would have to be informed first that his/her look-up behaviour was being recorded on an individual level. As a result, the procedure would no longer be unobtrusive. The same applies to analysing from which countries the users come or other meta-information associated with IP addresses. In our view, for most issues it is therefore not promising to examine log files at a granular level directed at an individual, but it is promising on a more quantitative level where individual look-up patterns are aggregated to large-scale datasets. This is exactly the nature of the log files supplied by the Wikimedia Foundation. The data used in this article consists of pre-aggregated log files without any IP-specific information of look-up behaviour (see section 2.2 of this article). Hence, different Wiktionary user groups (e.g., German native speakers vs. second language learners or article writers vs. article readers) or a sequence of look-ups from a single user cannot be analysed given the dataset at hand.

Which information users look up in dictionaries is primarily of interest for lexicographers whose aim is usually to satisfy the information needs of their users in the best way possible. It is also essential to know what users are interested in, when a new dictionary is being compiled (in order to choose the right headwords to be edited first). Furthermore, we think it is equally interesting for research into dictionary use in general to learn more about issues like the relationship between look-up frequency and corpus frequency, or the correlation between polysemy, or, more generally, the grade of ambiguity, and look-up frequency and other potentially interesting features that can be observed in log files aside from these relationships.

The following sections discuss these topics. The first part of the analysis will examine the correlation between look-up frequency and corpus frequency (section 3). This section summarizes previous findings (Koplenig et al. 2014). The second part of the analysis will consider questions of ambiguity and the role of (temporal) social relevance of words in connection with look-up frequency (section 4). The article ends with some concluding remarks. Before we

turn to the actual key findings, we must first clarify our understanding of the German Wiktionary and how our database is collected and pre-processed (section 2).

2. Our database

2.1 *The German Wiktionary*

Wiktionary is a multilingual dictionary which provides “two different approaches to encoding linguistic knowledge in multiple languages”. First, there are individual language editions for each language labelled with “the respective ISO 639 language code”. In the German Wiktionary, for instance, German is the describing language for the entries that is used for the graphical user interface and for labelling the individual items. Second, each Wiktionary edition normally “includes lexicon entries from multiple languages” (Meyer & Gurevych 2012: 263). Accordingly, the German edition of Wiktionary contains more than 350,000 headwords from 200 languages.³ This multilingual structure can lead to confusion. Fuertes-Olivera (Fuertes-Olivera 2009: 107–121), for instance, uses the term “Spanish Wiktionary” (2009: 112) when he refers to Spanish terms within the English-language edition and criticizes the domination of English without considering the actual Spanish-language edition (see also Meyer & Gurevych 2012: 264). When we use the term *German Wiktionary*, we refer only to the German headwords⁴ within the German-language edition of Wiktionary. This limitation (i.e., only the German headwords) is important for our purpose, to for instance combine log file data with corpus frequency lists.

Wiktionary is a crowd-sourcing project, created and edited by volunteers in a bottom-up process. Besides that, there is “the large-scale import of lexicon entries from copyright-expired dictionaries” (Meyer & Gurevych 2012: 262). The role of these automatic processes is often underestimated as Niederer and Van Dijk point out for Wikipedia:

“Although these researchers correctly observe significant changes in the ‘wisdom of crowds’ paradigm, they seem to be stuck in the antagonism of (few) experts versus (many) common users. Even if they notice the growing presence of non-human actors in the evolution of Wikipedia’s social dynamics, such as software tools and managerial protocols, they tend to underestimate their importance.”(2010: 1372)

With regard to lexicographic quality, this large scale import is often criticized because of the low quality of such contributions (e.g., Fuertes-Olivera 2009, Hanks 2012: 77–82). However, the quality of lexicographic data included in Wiktionary is not the focus here. In this article, we focus on observing

Wiktionary users. For this purpose, it is only crucial that the German Wiktionary is large enough and frequently used (on average 366,801 page counts per day during 2013) to have sufficient data to examine the aforementioned questions.

A last terminological clarification: The term “users of a dictionary” is usually reserved for the recipients of a dictionary, in contrast to lexicographers as authors. In terms of dictionaries with user-generated content (Lew 2013) this situation has changed. Meyer and Gurevych, for instance, use the term “users” mainly for the authors of Wiktionary, the “Wiktionarians” (2012: 271–272) and Lew points out that “the new model” [of user-generated content] “puts dictionary users in the shoes of lexicographers” which can be “aptly captured in the neologism *prosumer*, which is a blend of *producer* and *consumer*” (Lew 2013: 1). In contrast, if we speak of ‘observing Wiktionary *users*’, we only mean users who consult Wiktionary and “create” page counts, i.e. increase the number of visits for a particular page. With our database, we are not able to differentiate between users as authors and users as consumers, as already pointed out above.

2.2 Obtaining the data

The Wikimedia Foundation⁵ publishes log files for page view statistics in which all visits to any page are registered.⁶ The log files contain information about all projects of the Foundation (Wikipedia, Wiktionary and others) for a particular hour. Each row in a log file contains the name of the project (“de.d” for the German Wiktionary), the respective page’s name, the number of visits within the respective hour, and the size of the page’s content. Multiple page requests from the same IP address are treated as distinct page views. All downloads, selection processes, and statistical analyses were carried out using the software environment R (R Core Team, 2014). We aggregated the hourly log files for the German Wiktionary to daily, weekly, monthly, and one yearly dataset(s). In this article, mainly the dataset for the whole of 2013 is used. In Section 4.2, weekly and daily datasets are also introduced. To gather headword information (language, part-of-speech, ...), we also used a bzip2-compressed XML dump file⁷ of the current text and metadata of the pages in the German Wiktionary from March 12th, 2014. The dump was parsed with a custom R script which is available from the authors by request.

Previously (cf. Koplenig et al. 2014) we introduced the variable “searches per one million searches” (searches *poms*) that captures relative search frequency normed to one million searches. Because the Wiktionary log files do not contain search terms but only the number of visits within a specific time frame for specific pages, we will call the variable “visits in one million visits” here. Visits in one million visits were computed for all entries by multiplying the raw visits by the quotient of 1,000,000 and the sum of all visits.

Corpus frequency data was taken from a word form list based on the German Reference Corpus (DeReKo), representing a very large portion of the German language and “one of the major resources worldwide for the study of the German language” (Kupietz et al. 2010: 1848). The list contains frequency information for all spelt forms found in the German Reference Corpus.

Before presenting our analyses, we will investigate the effects of several reduction processes we applied to the dataset. It will be interesting to see how certain selection processes influence distribution profiles of parts-of-speech and frequency. To make our results comprehensible and reproducible for other researchers, we have to describe these different selection processes and the consequences on distributional profiles in our datasets in detail.

2.3 *Influences of selection processes*

As of March 12th 2014, the German Wiktionary contains 356,389 entries. As mentioned before, Wiktionary is a multilingual dictionary. So, there is also information regarding lemmas from other languages available in an edition of a specific language. 206,912 entries (58.1% of all entries) contain information for at least one German word.⁸ For 163,100 of the German entries (78.8%), we found the headword in the DeReKo frequency list and therefore have frequency information available. 70,891 (43.5%) of the entries with frequency information were accessed more than once in a million visits in 2013. These 70,891 entries are our primary database for subsequent analyses. We will call these entries selected entries. One might wonder whether the fact that only 34.3% ($70,891/206,912 \cdot 100$) of all German entries are analysed weakens the presented analyses. We argue that the first step is absolutely necessary for our analyses because we need to exclude headwords for which we cannot find any frequency information. We would further argue that the second step, in which we exclude headwords accessed less than once in a million visits in 2013, is actually a way to collect better data. We want to analyse those portions of Wiktionary that are important enough for the users to at least receive a minimum number of visits throughout a year. We chose a specific threshold (visited more than once in one million visits). This threshold can be ‘translated’ into raw numbers. In our case, an article has to be visited at least 134 times in 2013. On average, that means that it has to be visited 11.2 times a month or 0.37 times a day (or roughly once every 3 days) in order to be included in our analysis. In doing so, we focus on those parts of Wiktionary that ‘really matter’ in a large-scale analysis of user behaviour.

In Section 4.1, which contains the analysis of the influence of (non-)ambiguity on the number of visits per one million visits, we have to make an even narrower selection. There, we will exclude entries that have no information about the meaning of the described word (for further explanations see

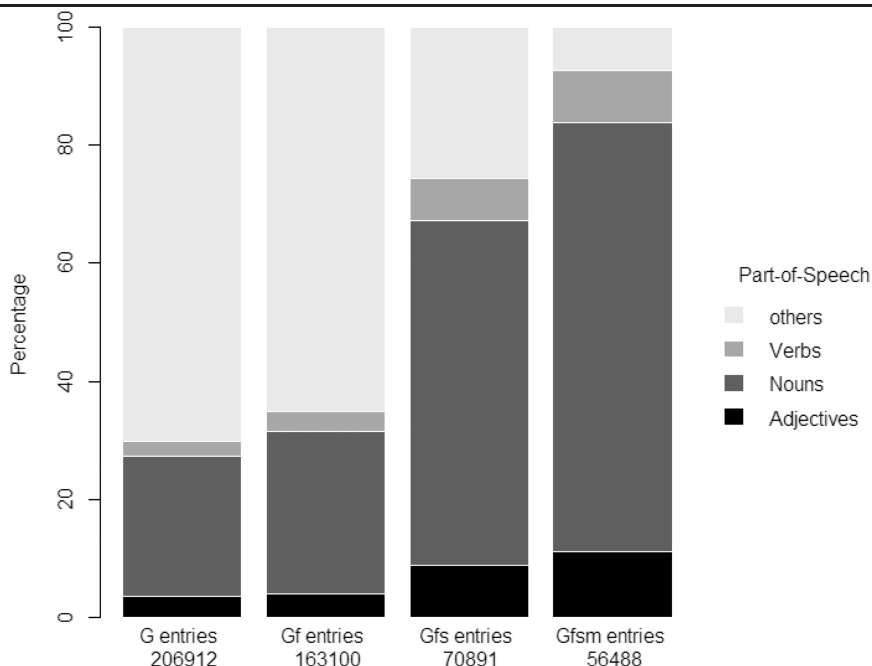


Figure 1: Part-of-speech distribution profiles of German entries (G), German entries with frequency information (Gf), selected German entries with frequency information (Gfs), and selected German entries with frequency and meaning information (Gfsm).

Section 4.1). This will narrow down our database to 56,488 words. Although we will use this subset only later in the article, we will include it in the current comparison. We will refer to it as selected entries with meaning information. Figure 1 gives an overview of the four datasets and the distribution profiles for parts-of-speech. Part-of-speech information was extracted from the Wiktionary entries. We chose to classify parts-of-speech by adjectives, nouns, verbs and other parts-of-speech. This seems to make particular sense if we look at the two right-most datasets, where adjectives, nouns, and verbs make up 74.3%, respectively, 92.6% of the overall database.

In Figure 1, it is clearly visible that the distribution profile of parts-of-speech changes over the different subsets of Wiktionary entries. The proportions of adjectives, nouns, and verbs do not change in relation to each other. However, the portion of other parts-of-speech declines considerably over the different stages of selection. This is also reflected in the raw counts (see Table 1). The main reason why so many “other” parts-of-speech are excluded is that inflected forms fall into the category “other”. Many of those inflected forms were excluded during our selection steps because they were not visited more than once in a million visits.

Table 1: Raw counts of different parts-of-speech for different Wiktionary datasets. See Figure 1 for a visualization of the data.

	Adjectives	Nouns	Verbs	other	total
German entries	7,226	49,186	5,456	145,044	206,912
& freq. inform.	6,653	44,884	5,177	106,386	163,100
& selected	6,319	41,359	4,996	18,217	70,891
& meaning inf.	6,272	41,063	4,976	4,177	56,488

2.4 *Closer inspection of German Wiktionary headwords with regard to corpus frequency bands and basic vocabulary*

To obtain a more detailed impression of the database we are dealing with in the current article, we are going to compare the Wiktionary dataset with a large-scale corpus of German on the one hand and a basic vocabulary list of German on the other hand. This section is a short version of a broader analysis we did (cf. http://www1.ids-mannheim.de/fileadmin/lexik/pdf-download/Working_Paper_Comparisons.pdf). For a comparison of Wiktionary and other resources like WordNet see Meyer and Gurevych (2012: 274–289).

A question in this context concerns the distribution of corpus frequencies: Which frequency bands are covered by Wiktionary headwords? To evaluate this question, we compared the Wiktionary dataset with large-scale corpus data, a DeReKo lemma list comprised of 326,949 German lemmas. It comes naturally to use a DeReKo lemma list as comparator because the corpus frequency measure for the Wiktionary headwords was taken from a spelt form list based on the same corpus. We used a lemma list (and not a spelt form list) for this comparison because, traditionally, lemmas are described in dictionaries. Note, that we still used word form frequencies from the spelt form list for later analyses (for instance concerning the influence of corpus frequency on look-up behavior) because we did not want to lose all the inflected forms contained in Wiktionary for these analyses.

The Wiktionary database contains 70,891 selected entries. Almost half (43.6%; 27,574 entries) of these inflected forms are not in the lemma list. This is not surprising because the lemma list, by definition, does not contain any inflected forms but Wiktionary does (and obviously they are looked up often enough to be included into our database). Another large group (40.3%) of headwords described in Wiktionary but not contained in the lemma list are nouns. These nouns comprise many geographic and proper names as well as terminological terms. Again, we refer to the working paper mentioned above for more details and examples. As a short summary we can say that the comparisons suggest that mainly mid- to highly frequent parts of the German

language are included in our Wiktionary database. Headwords in the lower frequency spectrum are either not described in Wiktionary or were de-selected because they were not visited often enough.

In contrast to the comparison with the big corpus list, we also want to take a closer look at the question whether the German Wiktionary contains the basic German vocabulary that a learner of German should know, i.e. a very small subset of German. For this purpose, we used a word list derived from the German part of the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). The CEFR is the common European basis for language curricula, exams, and textbooks for language learning. It classifies learners into six categories from A1 (“Basic Breakthrough User”) to C2 (“Proficient Mastery User”). B1 (“Threshold Independent User”) is the highest level for which the Goethe Institute⁹ provides an explicit word list (cf. Glaboniat et al. 2013). This list contains 3,405 words. Given this figure, we expect that all words should be described in the German Wiktionary. But this is not the case. 177 words (5.2% of all CEFR words) are not contained in the selected Wiktionary entries. No clear pattern is discernible for these missing words. They contain feminine occupational titles, numerals, and Anglicisms (Anglicisms are sometimes not marked as “German” in the German Wiktionary, so they were excluded from our dataset). Examples and further analyses are available in the working paper. For some of the missing items, there is no apparent reason why they should not be included in Wiktionary. A list of CEFR words missing in Wiktionary would provide a good mean to explore potential inconsistencies in a collaborative, crowd-sourced dictionary. However, this is not within the scope of the current article.

3. Correlation between corpus frequency and look-up frequency

If we want to compile a dictionary for a specific language, we have to make an important decision: Which words are to be included in the dictionary? In lexicographic projects, frequency of use which is measured using a corpus is often one of the main criteria when it comes to selecting headword candidates for a general-purpose dictionary (cf. De Schryver 2013b: 1385, Hanks 2012: 63). Only if the frequency of a word exceeds a defined threshold does it then become a candidate for inclusion in the dictionary. If more words exceed this frequency threshold than could be described appropriately in the dictionary given a limited amount of time and manpower, the threshold could just be raised accordingly. However, this means that it is implicitly assumed that it is somehow more important to include more frequent words instead of less frequent words. In this section, we want to investigate whether this is a good strategy. We already showed that dictionary users indeed look up frequent words (Koplenig et al. 2014). Here, we want to summarize our previous

Table 2: Distribution of 2013 log file data for the German Wiktionary

Category (visits per one million visits)	Percentage in Wiktionary log files 2013
1	65.44
2–10	26.75
11–49	6.70
50–500	1.10
500 +	0.01
<i>Total</i>	<i>100 (abs. 163,100)</i>

findings and evaluate them by transferring some of the methods we applied to the 2012 data from Wiktionary to the data from 2013.

Table 2 gives an impression of the distribution of the log file data. Visits per one million visits were rounded. The inherent Zipfian character can well be deduced from these figures. Almost two thirds (65.44%) of all entries were visited once per one million visits during 2013. A quarter (26.75%) of the entries was visited between two and ten times per one million visits. As a consequence, only 7.81% were visited 11 times or more per one million visits. This distributional profile is comparable to, e.g., word lengths. Words that are very short are highly frequent in language and make up the vast majority of all tokens within a language. Longer words are very infrequent in comparison to that – just like entries in Wiktionary that are visited very often. In Koplenig et al. (2014), we argued that log file and corpus frequency data pose a serious challenge to statistical techniques such as ordinary least squares (OLS) regression, Pearson’s or Spearman’s correlation. This is due to the non-linearity of the relationship between look-up and corpus frequency that cannot be simply solved by log-transforming the variables (also see O’Hara & Kotze 2010 on why this is seldom a good strategy). We are also faced with a very large number of rare events, which is typical for word frequency distributions (cf. Baayen 2008: 229). Rank-based techniques suffer from the fact that ranks in our observed distributions (for both look-up and corpus frequency) are far from being equidistant. Therefore, we employed something that we call a “simulation strategy”. The idea is to compile several “imaginary dictionaries” that contain a specific number of words that are selected due to criteria (e.g., corpus frequency) we can control in detail. These dictionaries can then be compared between one another in terms of look-up frequency.

To further analyse the log file data, we use the categories we already introduced in Koplenig et al. (2014). If an entry was visited at least once per one million visits during 2013, we say it was visited “regularly”. If it was visited at

Table 3: Proportion of regularly, frequently, and very frequently visited entries and their relationship to the number of included corpus frequency ranks

Included freq. ranks	% visited regularly	% visited frequently	% very frequently
10	100	100	100
50	100	100	100
100	100	100	87.0
500	100	99.2	73.6
1000	99.9	97.0	65.7
2000	99.3	92.4	58.1
5000	97.2	82.2	45.5
10000	92.1	73.4	36.5
20000	83.4	64.5	27.9
30000	77.6	59.0	22.8
50000	69.2	51.2	17.2

least twice per one million visits, we call it “frequently” visited. If it was visited at least 11 times per one million visits during 2013, we call it “very frequently” visited. Note, that the boundaries and names of these categories are chosen arbitrarily and merely have an illustrative function. It is also important to keep in mind that, by definition, categories are non-exclusive. Entries that are visited frequently are also visited regularly and entries that are visited very frequently are also visited frequently and regularly.

Using the “simulation strategy” briefly introduced above, we now create several dictionaries including more and more corpus frequency ranks based on the DEREKO corpus data. We then compare those dictionaries in terms of the proportion of entries visited regularly, frequently and very frequently. Table 3 and the corresponding visualization in Figure 2 clearly show the relationship between the number of included DEREKO ranks and the number of visits. The “imaginary dictionaries” are represented as tick marks on the x -axis in Figure 2 and the left-most column in Table 3. The more corpus frequency ranks are included in our dictionary, the less proportions of entries are visited regularly, frequently, and very frequently. If there was no relationship between corpus frequency and look-up frequency, we would expect these curves not to vary at all. This is clearly not the case. Also, there are systematic differences between the look-up categories. In the first two tiny dictionaries with the first 10 and 50 most frequent headwords described, all entries are visited very frequently (and therefore, by definition, frequently and regularly). If we consider the dictionary with the 2000 most frequent headwords, there are already

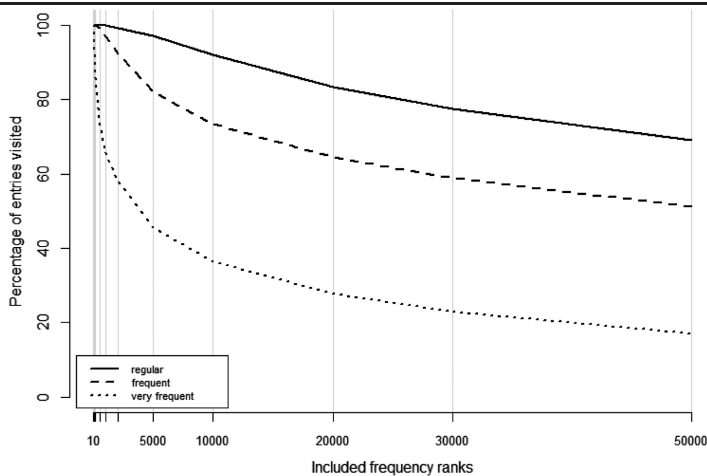


Figure 2: Percentage of entries visited regularly, frequently, and very frequently as a function of DEREKO rank.

considerable differences between those categories. Still, almost all entries are visited regularly given the Wiktionary log files of 2013. However, only 58% of all entries are visited very frequently. After a dictionary size of 500 entries, all categories diverge in terms of the proportions. Finally, in the largest dictionary, which contains the first 50,000 frequency ranks, almost 70% of the entries are still visited regularly but only roughly 17% are visited very frequently. So, obviously, there is a relationship between the number of corpus frequency ranks we include in our dictionary and the proportion of regularly, frequently, and very frequently visited entries.

In Koplenig et al. (2014: 242–245), we also showed how this relationship can be further carved out using additional information about the entries. We showed that the relationship becomes even clearer using word class information and a lemmatized word list. The latter was especially useful for log files of the “Digitales Wörterbuch der deutschen Sprache” (www.dwds.de).

As a short preliminary summary, we can thus state that frequency information based on a corpus can be used fruitfully for deciding which words to include in a general dictionary.

We now wish to concentrate on another question we also described briefly in Koplenig et al. (2014). It is the question of how “far down” in frequency ranks, frequency still matters in terms of look-up behaviour. De Schryver et al. state that “[c]orpus frequencies do not predict look-up behaviour beyond the top few thousand words of a language” (2006: 79). With the current dataset at hand, we can directly evaluate this statement for a large general online dictionary like the German version of Wiktionary.

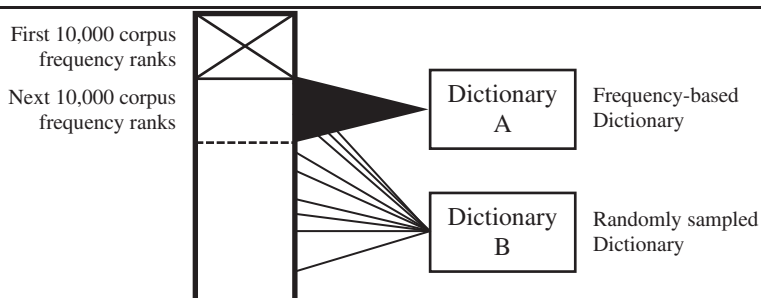


Figure 3: Simulation strategy for two competing dictionaries based on a larger data set where the first 10,000 corpus frequency ranks are excluded

Employing the simulation strategy described above, we can construct two contrastive dictionaries and compare look-up behaviour given the 2013 Wiktionary log files. The process is illustrated in Figure 3. In the first step, we exclude the first 10,000 entries in terms of corpus frequency. From the remaining entries, we build up two competing dictionaries. The first one (Dictionary A) comprises the entries with frequency ranks 10,001 to 20,000. For the second one (Dictionary B), we randomly sample 10,000 entries from the remaining entries. If look-up frequency also matters beyond the top 10,000 frequency ranks, we would expect Dictionary A to “perform better”. Here, performing better would mean that more entries are visited frequently and very frequently. The comparison of Dictionary A to B confirms our hypothesis. In Dictionary A, which contains the frequency ranks 10,001 to 20,000, 55.7% of the entries are visited frequently and 19.3% percent are visited very frequently. Dictionary B performs worse: 26.7% of the entries are visited frequently and 5.5% of the entries are visited very frequently. Given the analyses above, it is no surprise that the overall “success rate” is quite low for both dictionaries because the top 10,000 frequency ranks were excluded. But, obviously, the success rate of a frequency-based general dictionary is still better than of a randomly sampled one.

Figure 4 (see Table 4 for numerical data) shows how the relationship between a frequency-based and a randomly generated dictionary develops when more and more frequency ranks are excluded. Even when the first 30,000 entries are excluded (right-most group of bars in Figure 4), considerable differences can be observed between the two competing dictionaries. Here, 41.8% of the entries in a dictionary containing the next 10,000 frequency ranks are still visited frequently. In a randomly sampled dictionary, only around a quarter of the entries (23.0%) are visited frequently. Given this data, we conclude that a frequency-based dictionary “outperforms” a dictionary with randomly sampled entries – even when several thousand top frequency ranks are excluded. Therefore, we claim that frequency does matter – even in lower frequency bands.

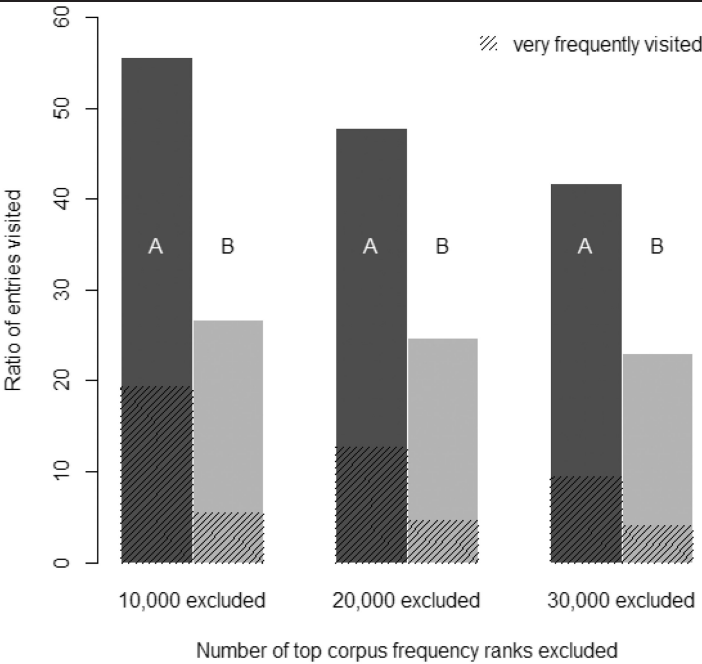


Figure 4: Ratios of entries visited for the frequency-based Dictionary A and randomly sampled Dictionary B with 10,000, 20,000 and 30,000 excluded frequency ranks. Bars represent frequently visited entries. Shaded areas represent very frequently visited entries.

Table 4: Comparisons of dictionaries based on corpus frequency and randomly sampled dictionaries.

Excluded corpus frequency ranks	Dictionary A: Next 10,000 frequency ranks	Dictionary B: 10,000 randomly sampled entries	Difference
first 10,000	frequent: 55.7%	frequent: 26.7%	frequent: 29
	very frequent: 19.3%	very frequent: 5.5%	very frequent: 13.8
first 20,000	frequent: 47.9%	frequent: 24.7%	frequent: 23.2
	very frequent: 12.7%	very frequent: 4.6%	very frequent: 8.1
first 30,000	frequent: 41.8%	frequent: 23.0%	frequent: 18.8
	very frequent: 9.5%	very frequent: 4.0%	very frequent: 5.5

One could now wonder how many frequency ranks have to be excluded until frequency really does not matter anymore. We would argue that this question cannot be answered given the available corpus data. Due to the Zipfian pattern of frequency distributions, corpora get less and less sensitive to frequency

differences in lower frequency ranges. Therefore, as soon as we enter very low regions of the frequency band, observed frequency differences get too small to show any effects on look-up frequency. Note that this does not have to be due to the fact that there really are no effects anymore – our available corpus data is simply not sensitive enough to capture them. One example could make that point clearer. Frequency rank 101 in our Wiktionary data is held by the word “seit” (English “since”). Its absolute DeReKo frequency is 2,873,373. If this word should take rank 100 (which is currently held by the word “Euro” with 2,934,200 occurrences), it would have needed 60,827 occurrences more. In contrast, the word on frequency rank 20,001, “Lifte” (“lifts”, “elevators”), is just 2 occurrences behind rank 20,000, “einhundert” (“one hundred”). So, with only 3 occurrences more, “Lifte” would hold the rank of “einhundert”. Given the total DeReKo counts of all Wiktionary headwords of over 2.9 billion, this difference can be considered random. So, discriminatory power is considerably degraded in lower frequency bands, even if we base our frequency ranks on a very large corpus of German like DeReKo.

To conclude this section, we reiterate our statement that corpus frequency indeed does matter for look-up frequency. The 2013 log files for the German Wiktionary support this statement as the 2012 data for Wiktionary and the DWDS dictionary already did (cf. Koplenig et al. 2014). However, we also want to emphasize that corpus frequency is no cure-all or “magic answer” (as was pointed out by De Schryver et al. 2006: 78–79) when it comes to compiling dictionaries. If the aim of a lexicographical project is to provide a general description of a language’s vocabulary, it simply is the best answer there is at the moment. If the aim is a specialized dictionary with a specific user group in mind, other criteria are relevant for selecting headword candidates (cf. e.g., Tarp 2008: 173–184, Granger & Paquot 2010, Bowker 2012, De Schryver & Prinsloo 2003).

4. Beyond frequency: Other effects on look-up frequency

In the previous sections, we showed that corpus frequency has an effect on how often a specific entry of the German Wiktionary is accessed. Although it is clear that there is a strong relationship between frequency of occurrence and look-up frequency, it is also quite clear that the former is not the *only* predictor of the latter. In the following sections, we will examine two other effects on look-up frequency.

4.1 *Mono- vs. polysemic words*

First, we want to investigate the influence of a semantic factor, namely if a lemma is mono- or polysemic. If a word is polysemic, there is potential uncertainty about the actual meaning of a word used in a specific context. One might

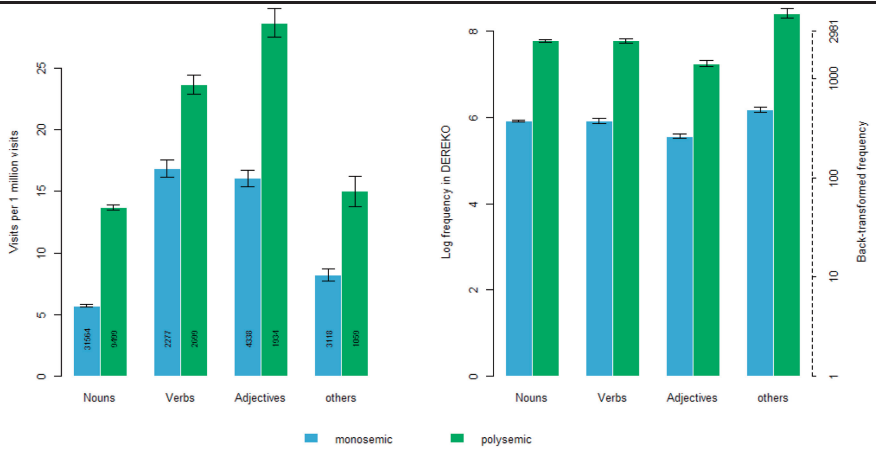


Figure 5: Visits per 1 million visits (left panel) and log DEREKO frequency (right panel) in relation to part-of-speech and (non-)ambiguity of the included lemmas. Numbers in the left panel bars denote the specific group’s size which is the same for the right panel. The numbers add up to 56,488 lemmas, the total of all included lemmas. The right axis in the right panel indicates raw (back-transformed) frequencies. Error bars indicate one standard error.

expect this uncertainty to lead to increased lookups of the word in a dictionary for language comprehension purposes. In language production tasks, one might also expect that words with multiple meanings are looked up more often, because more contextual information is needed to ascertain the correct meaning.

We analysed the XML dump of the German Wiktionary to extract information about the (non-)ambiguity of all lemmas. Lemmas were included into subsequent analyses if they received more than 1 visit in one million visits during 2013. We automatically extracted the information if a lemma has one or more meanings from the XML dump. Note that we did not use the *number* of listed meanings as a variable in our study because we are aware of the fact that lexicographic information provided in Wiktionary may not be reliable enough to rely on the exact number of listed meanings. Lemmas with no information regarding the number of meanings were excluded from the subsequent analysis.

Figure 5 (left panel) gives an overview of all 56,488 included lemmas, their part-of-speech, and if the lemma is mono- or polysemic. On the y axis, the mean visits per 1 million visits are plotted. At the bottom of the bars, the number of entries in this category is shown. The overall monosemic/polysemic ratio for our data is $41,297/15,191 = 2.72$. So, there are almost three times as many monosemic as polysemic entries in the analysed articles. This is in line with the ratio in the Digital Dictionary of the German Language (DWDS), an

academic dictionary written by lexicographic experts.¹⁰ In the DWDS, the monosemic/polysemic ratio is even higher with a value of 4.72.

As the graph shows, polysemic words are visited more often than monosemic words over all parts-of-speech, regardless of the specific group's size.

However, there is an important caveat when interpreting this relationship. As can be seen in the right panel of Figure 5, the corpus frequency of a word is highly correlated with (non-)ambiguity: Words that are lexically ambiguous are also more frequent. This holds for all parts-of-speech. As Gernsbacher notes for English words: “Printed frequency correlates strongly with multiplicity of meanings: The higher the probability of a given word appearing in printed English text, the more likely it has more than one meaning” (1984: 271). The same holds true for German.

If we combine this distribution with our previous findings, we come across a possible confound in our analysis: How can we tell whether the monosemy/polysemy contrast effect is reliable if monosemic words tend to be less frequent overall and we already know that less frequent words are looked up less frequently? In short: The effect of (non-)ambiguity on the number of visits per 1 million visits could be a ‘disguised’ frequency effect after all.

To tease apart these effects, we suggest a combined grouping and sampling strategy. First, we divide our datasets in log frequency deciles, i.e., we divide our dataset into 10 groups with a roughly equal group size over the whole frequency distribution (cf. OECD 2008: 131). To accomplish this, we have to expand or contract frequency intervals to “capture” an equal amount of lemmas in each group. This is the standard procedure to split a dataset into groups of equal size (most commonly is the median split, where datasets are divided into two groups with the median as division point). Figure 6 gives an impression of the frequency intervals we defined and how they relate to the frequency distribution as a whole. As expected, the right-most interval has to be fairly large in comparison to the others because there are few lemmas that are highly frequent. So, the group has to be expanded considerably in order to “capture” an equal amount of lemmas. Each group contains 5,650 lemmas with minor deviations¹¹.

In the second step, we go through each frequency group and take each polysemic lemma and match it with a random lemma from the monosemic

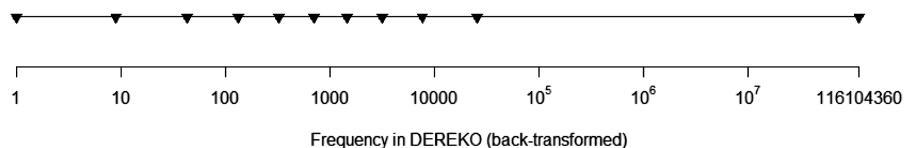


Figure 6: Frequency deciles determined on the basis of log frequencies. The black triangles denote the borders for each group. The axis is back-transformed from log to raw frequencies.

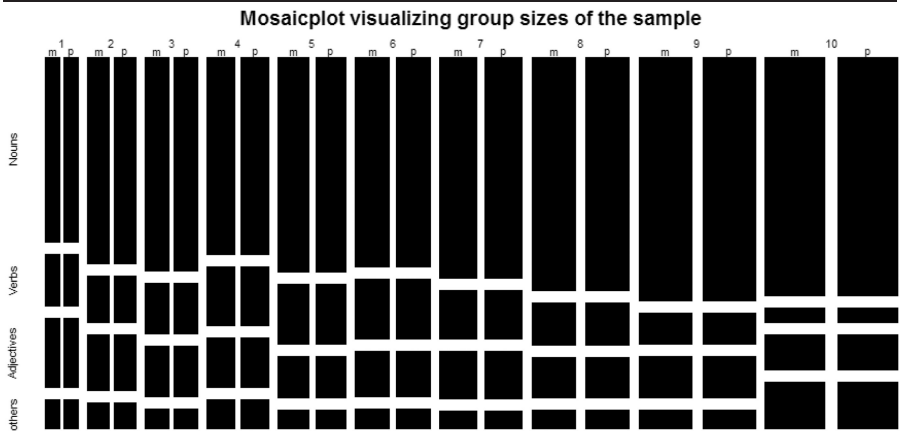


Figure 7: Group sizes in sampled dataset for frequency decile (numbers in columns), part-of-speech (rows) and (non-)ambiguity of lemmas (subdivision of columns; m = monosemic, p = polysemic). The sizes of the rectangles correspond to the number of cases within the given group.

group from the same part-of-speech group¹². This lemma pair then enters our sampled dataset which we will use for further analyses. The sampled dataset is therefore parallelized for (non-)ambiguity and part-of-speech in each frequency group. Figure 7 gives an overview of group sizes for the three relevant variables frequency group, part-of-speech, and (non-)ambiguity. Group sizes increase over the frequency band because higher frequency groups tend to contain more polysemic words (see the right panel of Figure 5). By employing this parallelization strategy, we made sure that the respective frequency/part-of-speech group always contains an equal amount of monosemic and polysemic words. Also, part-of-speech distribution within a specific frequency group is always proportional to the respective distribution in the overall dataset (note that in the highest frequency group 10, “other” words are contained more often because function words are highly frequent).

We can now analyse this parallelized dataset in regard to the effect of (non-)ambiguity. Since we still expect an effect of frequency, we include frequency group as a factor in our analysis. In Figure 8 (left panel), it can be seen how visits per 1 million visits tend to rise from lower to higher frequency groups for both monosemic and polysemic lemmas. For polysemic words, there seems to be kind of a “dent” for frequency groups 7 and 8. However, more importantly, one can clearly see that polysemic words are consistently visited more often than monosemic words. This holds true for all frequency groups. The difference also seems to get larger as frequency rises (which does not hold true for groups 7 to 9 but clearly, again, in the highest frequency group). To further investigate this effect, we conducted a simple linear model and predicted the

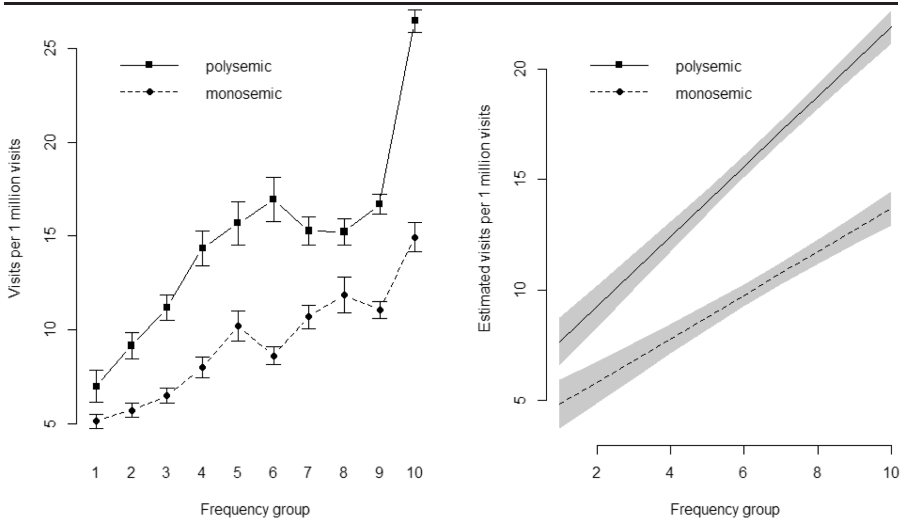


Figure 8: Comparison of (non-)ambiguity effect over frequency groups with associated standard errors (left panel) and interaction effect as estimated by a linear model predicting visits per 1 million visits by frequency group, (non-) ambiguity and the two-way interaction (right panel) with associated 95% confidence bands.

number of visits per one million visits by frequency group and the factor (non-) ambiguity as well as the interaction between the two factors. Both main effects are significant (frequency group: $\beta = 0.98$, $SE = 0.09$, $p = \min p$; (non-)ambiguity: $\beta = 2.23$, $SE = 0.91$, $p = 0.0142^{13}$). However, they are modulated by a significant two-way interaction ($\beta = 0.60$, $SE = 0.13$, $p = 4.43 \cdot 10^{-6}$). For this relevant interaction, we extracted the effect estimates from the linear model and plotted them along with associated 95% confidence bands in the right panel of Figure 8. The interaction has to be interpreted in a way that the mono- vs. polysemic contrast gets larger as the corpus frequency of the lemma increases. As the left panel of Figure 8 suggests, this interaction effect seems to be driven by frequency groups 1 to 6 and 10.

This interaction effect seems to be quite robust. The effect patterns remain the same for a variety of samples we took. We also made sure that the effects are not artefacts induced by excluding 35,691 lemmas which we did not find in DEREKO and therefore had no associated frequency measure. As an alternative, we added 1 to all raw frequencies¹⁴; therefore all non-matched lemmas receive a corpus frequency of 1 and can be included in the analysis. All effects remain stable when we use this alternative frequency measure. Another issue concerns the statistical method we applied for this analysis. We treated the frequency group as a continuous variable in the regression model. Of course, the grouping of lemmas into frequency deciles, which we needed for the

parallelization strategy described above, makes this variable an ordered factor, but not continuous in a strict sense (i.e., the difference between group 1 and group 2 could be different from the difference between group 8 and 9). However, the critical interaction effect remains stable if we include (ungrouped) log DEREKO frequency as a predictor in the model ($\beta = 0.77$, $SE = 0.12$, $p = 4.66 \cdot 10^{-10}$). Therefore, the interaction effect does not seem to be an artefact of the parallelization strategy we employed.

Given these analyses, we can conclude that polysemic words are indeed looked up consistently more frequently than monosemic words. This effect holds true for all frequency groups. Moreover, it seems to be more pronounced for lemmas with higher corpus frequencies than for lemmas which appear less often in natural language. Further analyses suggest that this interaction is especially driven by lemmas falling into the part-of-speech category “other” and by nouns. No interaction can be observed for verbs and adjectives. The fact that polysemic words are accessed more often in Wiktionary is not surprising: The fact that a word has several meanings often leads to more confusion which in turn brings participants of the speech community to look up the item in a dictionary. This can hold for speech recipients and producers. The significant interaction effect suggests that this effect of (non-)ambiguity is corroborated for items which are encountered more often in natural language. The interaction effect can be thought of as a superadditive effect of lemma frequency and polysemy on look-up behaviour.

4.2 *Temporary social relevance*

The main objective of this last section is to demonstrate a computationally cheap method of analysing the time series of look-up behaviour for entries in an electronic dictionary. Our concrete aim is to identify points in time where certain words are looked up extraordinarily often.¹⁵ To achieve this, we need to control for the overall trend of look-up frequency of each word. It is no surprise that look-up frequency varies over time. Words are looked up more or less often during the course of a year. This variation can be captured by the overall trend *within* a word’s visits. By controlling for these long-term trends, we also capture general look-up differences *between* words that stem, e.g., from the frequency effects outlined above. What we are currently interested in are rather short-term ‘peaks’ in the number of visits a specific word receives. The number of visits a specific word receives is the sum of the overall trend for this word and ‘noise’ which is not captured by this trend (cf. Beckett 2013: 92–95, 103, 109). This noise, or – informally speaking – what is left over after the overall trend has been considered, is exactly the kind of data we are interested in at this point. To extract this variable, we fitted a Tukey smoother using running medians of length 3¹⁶. This smoother captures the trend. The variable we are going to use in subsequent analyses is the difference between this

Table 5: The 10 entries of Wiktionary with the highest difference scores in 2013.

Headword	Date	Visits per 1 million visits	Difference score
Tribüne	May 6 th	286,188	286,176
fakultativ	January 18 th	98,064	96,254
Tribunal	May 6 th	67,516	67,496
Tribun	May 6 th	62,701	62,692
Grandezza	March 5 th	38,847	38,724
Komitee	August 30 th	17,277	17,217
reflektieren	June 14 th	16,680	16,596
Tribüne	May 7 th	31,140	15,397
Sommersonnenwende	June 21 st	10,747	10,716
Tribunat	May 6 th	10,684	10,676

smoother and the actual visits. We call this the difference score or residual visits. Using this technique, we can look beyond the effect of frequency and overall look-up tendencies of a specific word. In other words, this technique enables us to identify extraordinary look-up behaviour for individual words at individual points in time¹⁷. To extract interesting words, we rank words by their smooth-difference score. All highly ranking words have especially high proportions of unexplained variance in visits per one million visits in the respective week, day, or hour. Table 5 shows the top 10 difference scores of the German Wiktionary, the associated day in 2013 and the visits per one million visits the specific article received on that day.

With this method, we find – as expected – headwords which are topics of lexical discussion, like the word “Furor” (rank 17 on March 6th) which was part of a debate about sexism in Germany (Wolfer et al. 2014: 287). However, there are other noticeable words in certain periods of time, which are not directly related to discussions in society or politics that are lexical in nature. Figure 9 shows visits per 1 million visits per day for the entry “Borussia” (rank 32 on May 25th) during 2013. The line (in red online) in the left panel symbolizes the smoothed visits. The difference scores, i.e., the distances from the data points in the left panel to the smoothing line, are visualized in the right panel. “Borussia” is Latin for “Prussia” and part of the name of several German sports clubs. The most prominent ones are football clubs.

Peaks are identifiable in the difference scores for “Borussia” over time; symbolizing temporarily increased look-ups for “Borussia” in Wiktionary that cannot be explained by frequency of occurrence or overall search preferences alone. Each dashed vertical line in the right panel of Figure 9 represents one match in the knockout phase of the UEFA Champions League (CL)

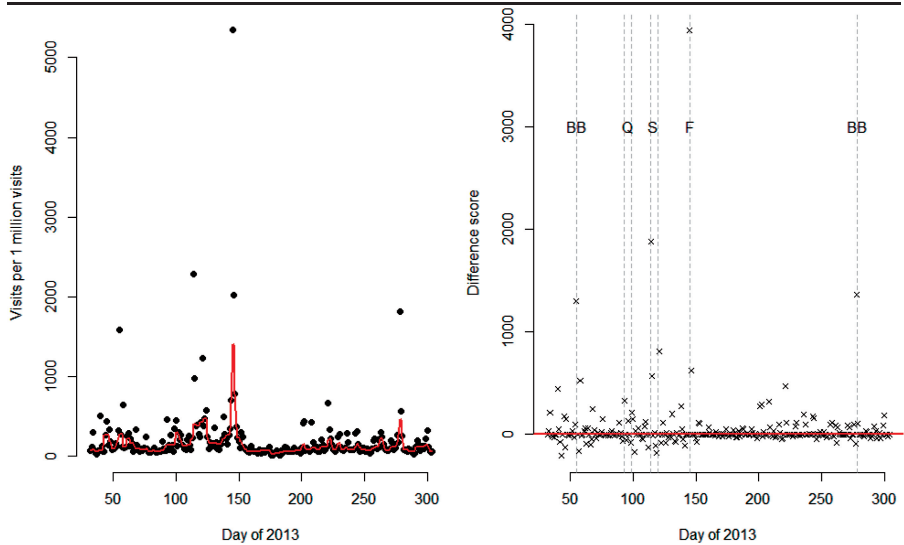


Figure 9: Visits per 1 million visits with smoothing line (left panel) and difference scores (right panel) for the entry “Borussia” from February until March 2013.

competition with the participation of Borussia Dortmund. Look-ups of “Borussia” sharply increase around match days. For the semi-finals (“S”) and especially the all-important final match (“F”), difference scores increase sharply around match days while the quarter finals (“Q”) do not trigger sharp increases. There are two other vertical lines (“BB”) which do not mark a match day in the CL. BB marks February 24th and October 5th, the days Borussia Dortmund competed against Borussia Mönchengladbach in the German first division. This match is associated with increased difference scores, too. In contrast, no other match in the German first division led to drastically increased difference scores for “Borussia”. Obviously, the popularity and importance of the CL competition led to repeated increases in the social relevance of the term “Borussia”. Also, when both Borussias competed against each other in the national championship, public interest in the somewhat cryptic name part also increased. In comparison to the “Furor” case presented above, the look-up behaviour concerning “Borussia” is more surprising. There is no lexico-semantic debate involved that could persuade people to look up “Borussia”. Increased media coverage and general public awareness concerning a football club alone seems sufficient to trigger noticeable increases in look-up behaviour. Another example is the word “larmoyant” (English “lachrymose” meaning “tearfully sentimental”) which was used in a sports commentary in a friendly match between the French and German national football teams. Here, the commentator described one specific German national as being too

“larmoyant” which led to sharply increased lookups within the same hour (which is the minimum temporal resolution available for the Wiktionary log files).

There are several more interesting cases extractable from the Wiktionary log files that we cannot report here. Social relevance in other cases was induced by a variety of social contexts like TV game shows (“Tribüne” and all related words, see ranks 1, 3, 4, 8, and 10 in Table 5) and even astronomical events like a solstice (“Sommersonnenwende”, see rank 9 in Table 5). The social context (in a very broad sense) directly influences look-up frequencies in internet dictionaries. Most importantly, however, the methodology for identifying “unusual” peaks in look-up behaviour on the basis of the difference scores is widely applicable. Also, the timeframe of the smoothing process (i.e., whether one uses hours, days or weeks as the unit for smoothing) is adaptable. Social relevance is just one (rather obvious) factor that can be investigated using this method.

5. Concluding remarks

Log files are sensitive data that commercial publishers would never publish, since they are part of their business secrets. The Wikimedia products are therefore a valuable resource for research, since everyone can evaluate these resources for free and without copyright constraints. Generally, the analysis of log files does not permit any conclusions on an individual level if data privacy issues are taken seriously. We believe that this should always be the case. But also without this level of granularity, we hope that we have demonstrated that quantitative evaluations of log files can give profound insights into general patterns of look-up behaviour. Those evaluations can give “solid empirical evidence” (Lew 2011b: 3) for questions such as the most searched words in a dictionary or the connection between corpus frequency or grade of ambiguity and look-up frequency. In addition, social events and/or related linguistic discussions in social discourse can be observed by the analysis of log files, beyond the intuitively expectable extent. This makes it possible to draw conclusions on linguistic reflection, especially on the temporal relations between social events and look-up acts in dictionaries.

These findings contain no magic answer. We doubt that there is such a thing and even doubt that it is the task of science to find one. Rather, these results are solid empirically examined pieces that contribute a small part to the phenomenon ‘dictionary use’, or, as indicated, to the topic of how thinking about language is reflected in a collaborative dictionary like Wiktionary.

In summary, we can state that observing Wiktionary users is a multifaceted task. The results of this research task should not be overestimated, as we think there is still a lot of not yet fully exploited potential one can work on.

1 We would like to thank our colleague Peter Meyer for supporting us in this research with technical assistance and many fruitful discussions. We would also like to thank our three anonymous reviewers for some very valuable remarks which helped us to improve our article.

2 We know that it is common practice to analyse ‘who does what online’ – as one anonymous reviewer remarked – without taking into account the relevant provisions concerning the protection of privacy. We think as researchers financed by the public, we are to respect these legal constraints.

3 “361.910 deutschsprachige Einträge zu über 200 Sprachen” (last accessed April 1st 2014).

4 Headwords labelled with “(Deutsch)”, for instance “Haus (Deutsch)” <http://de.wiktionary.org/wiki/Haus> (last accessed April 1st 2014).

5 Cf. <http://wikimediafoundation.org> (last accessed August 20th 2014).

6 The base URL for obtaining log files from 2013 is <http://dumps.wikimedia.org/other/pagecounts-raw/2013/> (last accessed August 20th 2014).

7 A complete list of XML dumps from this date can be accessed at <http://dumps.wikimedia.org/dewiktionary/20140312/> (last accessed August 20th 2014).

8 Hence, there are 149,477 entries (41.9%) for non-German words in the German Wiktionary. Those entries were excluded from subsequent analyses. Some entries contain information regarding a German word and a word from another language, e.g., the entry for “last”.

9 The Goethe Institute is the institution responsible for CEFR-related activities in Germany.

10 We are grateful to Axel Herold who advised us with the figures for the DWDS. And we thank one anonymous reviewer for the recommendation to check this ratio against other resources in order to rule out that it is not something special in the Wiktionary data.

11 Deviations stem from the fact that the dataset contains 56,488 cases, which is not divisible by 10 without remainder. Also, cases which lie exactly on a group border are assigned to the group which is right of the border.

12 Whenever there were fewer monosemic than polysemic lemmas for a certain part-of-speech in a certain frequency group, we included pairs as long as there were still monosemic lemmas available to match the polysemic lemmas. After that point, we stopped including pairs for this group. This happened for all parts-of-speech in frequency decile 10 and for verbs in the deciles 6 through 10.

13 Whenever the level of significance reaches the minimal value representable by R (which is $p = 2.2 \cdot 10^{-16}$) we use the notation $p = \min p$. This means that the probability of an error is virtually 0.

14 We are aware that this is considered a suboptimal strategy for dealing with zero frequencies (cf. Brysbaert & Diependaele 2013). Here, we only use this strategy to test if the overall results change when lemmas with zero frequency are included. In this light, the exact strategy used to deal with zero frequencies is not relevant.

15 Results of this study are also reported in Wolfer et al. (2014: 286–289).

16 To do this, we used the default behaviour of the function `smooth()` provided by the ‘stats’ package of the statistical programming language R.

17 Of course, these differences can also take negative values. Indeed, many of them do. This means that a word was visited less often in a particular week than would be expected given the word’s overall trend.

References

- Baayen, R. H. 2008.** *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.
- Beckett, S. 2013.** *Introduction to Time Series Using Stata*. College Station: Stata Press.
- Bergenholtz, H. and M. Johnsen. 2005.** Log Files as a Tool for Improving Internet Dictionaries. *Hermes. Journal of Language and Communication Studies*, 34: 117–141.
- Bergenholtz, H. and M. Johnsen. 2013.** User Research in the Field of Electronic Dictionaries: Methods, First Results, Proposals. In R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds), *Dictionaries. An International Encyclopedia of Lexicography Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, Boston: De Gruyter, 556–568.
- Bowker, L. 2012.** Meeting the Needs of Translators in the Age of E-Lexicography: Exploring the Possibilities. In S. Granger and M. Paquot (eds), *Electronic Lexicography*. Oxford: Oxford University Press, 379–397.
- Brysbaert, M. and K. Diependaele. 2013.** Dealing with Zero Word Frequencies: A Review of the Existing Rules of Thumb and a Suggestion for an Evidence-Based Choice. *Behavior Research Methods*, 45.2: 422–430.
- De Schryver, G.-M. 2013a.** The Concept of Simultaneous Feedback. In R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, Boston: De Gruyter, 548–556.
- De Schryver, G.-M. 2013b.** Tools to Support the Design of a Macrostructure. In R. H. Gouws, U. Heid, W. Schweickard and H. E. Wiegand (eds), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, Boston: De Gruyter, 1384–1395.
- De Schryver, G.-M. and D. Joffe. 2004.** On How Electronic Dictionaries Are Really Used. In G. Williams and S. Vessier (eds), *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Université de Bretagne Sud, 187–196.
- De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert. 2006.** Do dictionary users really look up frequent words? - on the overestimation of the value of corpus-based lexicography. *Lexikos*, 16: 67–83.
- De Schryver, G.-M. and D. J. Prinsloo. 2003.** Compiling a Lemma-Sign List for a Specific Target User Group: The Junior Dictionary as a Case in Point. *Dictionaries: Journal of The Dictionary Society of North America*, 24: 28–58.
- Fuertes-Olivera, P. A. 2009.** The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary. In H. Bergenholtz, S. Nielsen and S. Tarp (eds), *Lexicography at a crossroads: dictionaries and encyclopedias today, lexicographical tools tomorrow* (Linguistic Insights - Studies in Language and Communications). Bern et al.: Peter Lang, 99–134.
- Gernsbacher, M.A. 1984.** Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology / General*, 113/2: 256–281.
- Glaboniat, M., M. Perlmann-Balme and T. Studer. 2013.** *Zertifikat B1. Deutschprüfung für Jugendliche und Erwachsene. Prüfungsziele, Testbeschreibung*. Ismaning: Hueber.

- Granger, S. and M. Paquot. 2010.** The Louvain EAP Dictionary. In A. Dykstra and T. Schoonheim (eds), *Proceedings of the XIV EURALEX International Congress*. Leeuwarden/Ljouwert: Fryske Akademy, 321–326.
- Hanks, P. 2012.** Corpus Evidence and Electronic Lexicography. In S. Granger and M. Paquot (eds), *Electronic lexicography*. Oxford: Oxford University Press, 57–82.
- Hult, A.-K. 2012.** Old and New User Study Methods Combined – Linking Web Questionnaires with Log Files from the Swedish Lexin Dictionary. In R. V. Fjeld and J. M. Torjusen (eds), *Proceedings of the 15th EURALEX International Congress 2012*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 922–928.
- Koplenig, A., P. Meyer and C. Müller-Spitzer. 2014.** Dictionary Users Do Look up Frequent Words. A Log File Analysis. In C. Müller-Spitzer (ed.), *Using Online Dictionaries*. Berlin, Boston: De Gruyter, 229–249.
- Kupietz, M., C. Belica, H. Keibel and A. Witt. 2010.** The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, D. Tapias, M. Rosner, S. Piperidis, J. Odijk, J. Mariani, and K. Choukri (eds), *Proceedings of the 7th conference on International Language Resources and Evaluation*. (LREC-10). Valetta, Malta: European Language Resources Association (ELRA), 1848–1854.
- Lew, R. 2011a.** User Studies: Opportunities and Limitations. In K. Akasu and U. Satoru (eds), *ASIALEX2011 Proceedings Lexicography: Theoretical and practical perspectives*. Kyoto: Asian Association for Lexicography, 7–16.
- Lew, R. 2011b.** Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography*, 24.1: 1–4.
- Lew, R. 2013.** User-Generated Content (UGC) in English Online Dictionaries. In A. Abel and A. Klosa (eds), *Ihr Beitrag bitte! - Der Nutzerbeitrag im Wörterbuchprozess*. (OPAL - Online Publierte Arbeiten Zur Linguistik). Mannheim: Institut für Deutsche Sprache, 9–30.
- Meyer, C. M. and I. Gurevych. 2012.** Wiktionary: A New Rival for Expert-Built Lexicons? Exploring the Possibilities of Collaborative Lexicography. In S. In Granger and M. Paquot (eds), *Electronic Lexicography*. Oxford: Oxford University Press, 259–291.
- Niederer, S. and J. van Dijck. 2010.** Wisdom of the Crowd or Technicity of Content? Wikipedia as a Sociotechnical System. *New Media & Society*, 12.8: 1368–1387.
- OECD. 2008.** OECD Glossary of Statistical Terms, Paris: Organisation for Economic Co-operation and Development. <http://www.oecd-ilibrary.org/content/book/9789264055087-en> (November 18th, 2014).
- O'Hara, R. B. and D. J. Kotze. 2010.** Do Not Log-Transform Count Data. *Methods in Ecology and Evolution*, 1.2: 118–112.
- R Core Team. 2014.** R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/> (November 18th, 2014).
- Santos, D. and A. Frankenberg-Garcia. 2007.** The Corpus, Its Users and Their Needs: A User-Oriented Evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12.3: 335–374.
- Tarp, S. 2008.** *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer Verlag.
- Tarp, S. 2009.** Reflections on Lexicographical User Research. *Lexikos*, 19: 275–296.

- Tiberius, C. and J. Niestadt. 2014.** Dictionary Use: A Case Study of the ANW Dictionary. In C. Tiberius and C. Müller-Spitzer (eds), *Dictionary Use: A Case Study of the ANW Dictionary. Research into dictionary use - Wörter-buchbenutzungsforschung*. 5. Arbeitsbericht des wissenschaftlichen Netzwerks “Internetlexikografie”, (OPAL - Online Publierte Arbeiten Zur Linguistik). Mannheim: Institut für Deutsche Sprache, 27–33. <http://multimedia.ids-mannheim.de/mediawiki/web/images/7/7f/Preprint-V1.pdf> (June 12th, 2014).
- Wolfer, S., A. Koplenig, P. Meyer and C. Müller-Spitzer. 2014.** Dictionary Users Do Look up Frequent and Socially Relevant Words. Two Log File Analyses. In A. Abel, C. Vettori and N. Ralli (eds), *Proceedings of the XVI Euralex International Congress: The User in Focus*. Bolzano: EURAC research, 281–290.